Identification strength with a large number of moments

Hyojin Han* Eric Renault†

April 3, 2020

Abstract

This paper studies how identification is affected in GMM estimation as the number of moment conditions increases. We develop a general asymptotic theory extending the set up of of Chao and Swanson (2005) and Antoine and Renault (2009, 2012) to the case where moment conditions have heterogeneous identification strengths and the number of them may diverge to infinity with the sample size. We also allow the models to be locally misspecified and examine how the asymptotic theory is affected by the degree of misspecification. The theory encompasses many cases including GMM models with many moments (Han and Phillips (2006)), partially linear models, and local GMM via kernel smoothing with a large number of conditional moment restrictions. We provide an understanding of the benefits of a large number of moments that compensate the weakness of individual moments by explicitly showing how an increasing number of moments improves the rate of convergence in GMM.

JEL classification: C01, C13, C14

Keywords: Generalized Method of Moments, Weak Identification, Alternative Asymptotic Theory

*Gregory and Paula Chow Center for Economic Research (GCCER) and Department of Statistics, School of Economics, Xiamen University, hyojin_han@xmu.edu.cn

†Department of Economics, University of Warwick, Eric.Renault@warwick.ac.uk

1 Introduction

Since the seminal papers of Chao and Swanson (2005) and Han and Phillips (2006), the asymptotic theory of the Generalized Method of Moments (GMM) of Hansen (1982) has been extended to accommodate the presence of weak identification and large numbers of moment conditions.

We provide in this paper a general setup for deriving a feasible asymptotic theory for a GMM estimator in the context of many moment asymptotics with heterogeneous identification strengths. This setup and results can be seen as an extension of the work of Chao and Swanson (2005), Han and Phillips (2006) as well as of Antoine and Renault (2009, 2012). We extend the 2SLS results (not the LIML ones) of Chao and Swanson (2005) (resp. the nonlinear GMM results of Han and Phillips (2006)) by considering nonlinear moments (resp. moments with heterogeneous identification strengths) while we extend the work of Antoine and Renault (2012) by allowing the number of moment conditions to go to infinity with the sample size. As it has been shown by the nonlinear extension of Staiger and Stock (1997) in Stock and Wright (2000), nonlinearity is not innocuous in the context of weak instrument asymptotics. We need to maintain a uniform boundedness assumption about the empirical process of sample moments to prove a consistency result for the GMM estimator in the nonlinear case. In this respect, we also contribute to extend the Stock and Wright (2000)'s approach to a case where the number of moments goes to infinity and instruments have heterogeneous identification strengths.

While both Chao and Swanson (2005) and Hausman et al. (2012) have also tackled the case of large numbers of instruments with heterogeneous identification strengths, Newey and Windmeijer (2009) have arguably proposed the most general framework to deal with many weak moment conditions. They show that the continuously updated GMM estimator (CUE) would be consistent and asymptotically normal, even in some situations where moment conditions are weaker than the ones considered in the present paper, but with an extra correction term in the asymptotic variance matrix.

Our goal in this paper is different. We rather consider GMM with a deterministic sequence of weighting matrices under identification conditions that allow us to end up with the "textbook asymptotic variance matrix," following Newey and Windmeijer (2009)'s terminology. Note that by eliciting a deterministic weighting matrix, we overlook not only the advantages of CUE (as enhanced by Newey and Windmeijer (2009)) but also the complications due to preliminary estimation of a large variance matrix, as discussed in Donald el al. (2003). Moreover, as Han and Phillips (2006) (see their corollary 14 and subsequent comments about the case of weak moment conditions), we acknowledge that "the moment conditions should be stronger than in the Stock and Wright (2000) setting". Our main goal is precisely to allow a researcher to perform GMM inference as usual, as long as our conditions of identification strength of moment conditions are fulfilled. These conditions are actually testable (see Antoine and Renault (2019)) and, when moment conditions are heteroskedastic

or nonlinear, lead to a novel "rule-of-thumb" to check that standard GMM inference is reliable (see Frazier et al. (2019)).

It is precisely because our goal is to state the conditions of validity of standard GMM inference, that the extensions of Chao and Swanson (2005) (CS hereafter) and Han and Phillips (2006) (HP hereafter) that we propose are non-trivial and potentially useful. The key is to realize what one should understand by the aforementioned "textbook asymptotic variance matrix". If one believes that any parameter (or any linear combinations of these parameters) has an estimator whose asymptotic standard error, properly scaled by the square root of an effective sample size, that looks like the textbook formula, it is plainly wrong. The asymptotic variance of the GMM estimator looks like the textbook one only after a convenient change of basis in the parameter space while the effective sample sizes differ depending on the different directions in this new basis. The beauty of the asymptotic theory derived by Antoine and Renault (2012) (AR12 hereafter), (see also Cheng (2015)) is that, due to Studentization, the textbook formulas for Wald inference based on GMM estimators can be safely applied, even though without a preliminary change of basis (and proper rescaling), GMM estimators are not well behaved. While in AR12, the rescaling was based on an index of identification strength (index between 1 for weak identification and square root of T for strong identification) we show in the present paper that the effective sample size (a generalized "concentration parameter") is now the product of this squared index of strength and the number of moment conditions with the same strength index.

While, as explained above, our asymptotic theory should help the empirical researcher for a variety of contexts with many weak instruments, we show that it may also help in other econometric contexts which, albeit not generally described as "weak identification," share a similar mathematical structure. Moreover, they justify the additional extension that we propose, namely allowing for some degree of local misspecification. More precisely, we revisit the standard concept of local misspecification (see e.g. Hall (2005)) by considering biased moments whose bias decreases with the sample size, but possibly slower than square root of T when the identification index is indeed slower. In other words, a side benefit of (nearly) weak identification is to allow the use of more severely biased moment conditions since, with weaker identification, the bias is less swiftly contaminating the parameter estimation. We put forward two important examples of this situation.

First, we revisit the case of (partially) linear models with many regressors, as studied by Anatolyev (2012) and Cattaneo et al. (2018). It paves the way for a new view about moment weakness, since our focus of interest is weakness that is not due to the poor correlation but due to the asymptotic spanning of the partial regression of interest by a possibly infinite number of regressors. Moreover, the fact that in practice we always use only a finite number of regressors implies that estimators of the regression of interest are biased in finite sample (omitted variable bias), although this bias vanishes asymptotically when increasing the number of observations allows the researcher to increase

accordingly the number of regressors. This phenomenon is a motivation for our more general setup where, on top of weak identification, local misspecification is allowed.

Second, we also revisit the local GMM approach introduced by Lewbel (2007) and Gagliardini et al. (2011). It is typically a case where our general framework of heterogeneity of rates of identification/convergence is well suited, because different degrees of smoothness of the regression functions of interest may allow kernels of different orders to estimate them. We have in particular in mind several asset pricing applications. Asset prices or returns are often characterized through conditional moment restrictions conditioning on some lagged variables. Then the lack of predictability of some asset returns from some lagged variables may lead to the weak instrument problems. In addition, the standard GMM that exploits unconditional moment restrictions deduced from those conditional moments requires the assumption that the conditional moments are available for any value of the conditioning variables. However, we may need to consider the cases where moments are valid only for a given value of conditioning variables (local moment restrictions). For example, only a few options are traded on a daily basis so the prices of options in general are observed at only one given date or a few dates (see e.g., Gagliardini et al. (2011)). The local GMM approach by Lewbel (2007) and Gagliardini et al. (2011) addresses this issue through kernel estimation of local moment conditions. Then the several rates of identification may be present simultaneously due to the different bandwidths for kernel estimation of different local moment conditions. Moreover, we may need to estimate a large number of such local moment conditions as a large number of cross sections of the option prices may be available. In these examples, the poorer information from the subsets of moment conditions with weaker identification strength may contaminate the stronger information from the other sets of moments regarding the estimation. This problem necessitates a general model to accommodate such heterogeneity so that we can optimally exploit different qualities of information. We provide in this paper a general set up for stating a feasible asymptotic theory for a GMM estimator in the context of many moment asymptotics with heterogeneous identification strengths.

The rest of the paper is organized as follows. Section 2 discusses asymptotic identification and consistency under many moment conditions through a general theory and a discussion of the three classes of motivating examples as well. Section 3 shows the asymptotic normality of GMM estimators with possibly heterogeneous rates of convergence. Section 4 concludes. Most of the proofs are relegated to the supplemental appendix. The appendix also provides the discussion of the linear IV model by CS and the illustration of the various possibilities of the asymptotic normality with different convergence rates through the three examples introduced in section 2.

2 Moment Conditions and Consistent GMM Estimation

2.1 Identification strength

As already announced, our general framework extends the one proposed by HP in order to explicitly take into account, as in AR12, the possibility of heterogeneity of identification strengths across different moment conditions. Like HP, we consider a set of sample moments computed from a double array of moment functions. More precisely, if $\psi_{t,T}(\theta)$ is a J_T-dimensional random vector, function of a vector of p unknown parameters, the sample mean of interest is:

$$\bar{\psi}_T(\theta) = \frac{1}{T}\sum_{t=1}^{T}\psi_{t,T}(\theta).$$

HP argue that, in order to measure the total identification strength of an increasing number J_T of possibly weak moment conditions, it is worth introducing a real sequence $c_T, T = 1, 2, ...,$ such that:

$$\lim_{T\to\infty}\frac{1}{c_T}\left\|E\left[\bar{\psi}_T(\theta)\right]\right\|^2 = \gamma\left(\theta\right). \tag{2.1}$$

However, this argument needs to be revisited to take into account the heterogeneity of identification strengths. To see that, we adopt a formalism similar to AR12 by writing:

$$E\left[\bar{\psi}_T(\theta)\right] = \frac{\Lambda_T}{\sqrt{T}}\rho_T(\theta)$$

where $\rho_T(\theta)$ stands for a bounded deterministic sequence of J_T-dimensional functions. The squared matrices $\Lambda_T, T = 1, 2, ..,$ of dimension J_T , and assumed to be nonsingular, characterize the way the total identification strength is spread across the various moment conditions. Even though this assumption may be too restrictive in specific examples (see e.g. Frazier et al. (2019) for a Probit model with simultaneity), it is worth interpreting the sequence of matrices Λ_T when they are diagonal. With the terminology of HP, the diagonal coefficients $\lambda_{j,T}, j = 1, ..., J_T$ "embody the weakness of the signal strength in terms of the individual moment functions".

Example 1: GMM with Many Moment Conditions and Diagonal Λ_T
For any $\theta \in \Theta$ suth that $\rho_T(\theta) \neq 0$, we can write:

$$\left\|E\left[\bar{\psi}_T(\theta)\right]\right\|^2 = \frac{J_T\bar{\lambda}_T^2(\theta)}{T}\frac{\left\|\rho_T(\theta)\right\|^2}{J_T} \tag{2.2}$$

where:

$$\bar{\lambda}_T^2(\theta) = \frac{\sum_{j=1}^{J_T} \rho_{j,T}^2(\theta)\lambda_{j,T}^2}{\sum_{j=1}^{J_T} \rho_{j,T}^2(\theta)}$$

is a weighted average of the diagonal coefficients $\lambda_{j,T}, j = 1, ..., J_T$ with respective weights $\rho_{j,T}^2(\theta)$.

The weakness of the signal strength in terms of the individual moment functions $\psi_{j,t,T}(\theta), j = 1, .., J_T$ is characterized by the diagonal terms $\lambda_{j,T}, j = 1, ..., J_T$ and $\bar{\lambda}_T(\theta)$ characterizes a kind of average identification strength. It is natural to assume (see assumption 1 below for a formal statement) that we have the following uniform limit on Θ:

$$\lim_{T \longrightarrow \infty} \frac{\|\rho_T(\theta)\|^2}{J_T} = \delta(\theta). \tag{2.3}$$

By comparing (2.1) and (2.2), we see that with $\gamma(\theta) = \delta(\theta)$

$$c_T = c_T(\theta) = \frac{J_T \bar{\lambda}_T^2(\theta)}{T}.$$

In particular, if the matrix Λ_T is scalar ($\Lambda_T = \lambda_T Id$), so that $\bar{\lambda}_T^2(\theta) = \lambda_T^2$, the product $J_T\lambda_T^2$ may be interpreted as a general version of concentration parameter (see CS for the case of a genuine concentration parameter in a linear IV model) with the three cases of interest:

(i) $\lambda_T = \sqrt{T}$ for strong identification,

(ii) $\lambda_T = 1$ for weak identification a la Staiger and Stock (1997) and Stock and Wright (2000),

(iii) $\lambda_T = T^\gamma, 0 < \gamma < 1/2$, for nearly-weak (or nearly-strong) identification a la Hahn and Kuersteiner (2002), Caner (2008), Antoine and Renault (2009, 2012), as well as Andrews and Cheng (2012).

We will argue in the next subsection that the HP's interpretation of c_T as a measure of the total identification strength is ambiguous in case of heterogeneity of identification strengths. We first note that in the heterogeneous case, this measure will in general depend on θ as the average identification strength (or generalized concentration coefficient) $J_T\bar{\lambda}_T^2(\theta)$. Therefore, while the main identification assumption of HP (see their assumption 2(ii)) is

$$\lim_{T \longrightarrow \infty} \frac{J_T}{T c_T(\theta)} = \lim_{T \longrightarrow \infty} \frac{1}{\bar{\lambda}_T^2(\theta)} = \alpha(\theta) \in [0, +\infty),$$

we need to discuss values of θ for which this assumption is relevant. We will show in the next subsection that these values must be discussed in relation to the family of GMM estimators of

interest. Note however that we will always maintain the other part of assumption 2(ii) in HP:

$$\lim_{T\longrightarrow\infty} Tc_T(\theta) = +\infty.$$

2.2 GMM estimators of interest

We will be interested hereafter in the properties of a GMM estimator $\hat{\theta}_T$ of θ defined as follows:

Definition 1: A GMM estimator $\hat{\theta}_T$ of θ based on the moment functions $\bar{\psi}_T(\theta)$ is a solution of:

$$\hat{\theta}_T = \arg\min_{\theta} \bar{\psi}_T(\theta)' W_T \bar{\psi}_T(\theta)$$

where W_T is a sequence of (possibly random) positive definite weighting matrices such that:

$$\|W_T\| = O_P(1), \left\|W_T^{-1}\right\| = O_P(1).$$

Note that for a symmetric positive semi-definite matrix W_T of dimension J_T, we will use the spectral norm $\|W_T\|$, that is the largest eigenvalue of W_T. In other words, eigenvalues of the sequence of weighting matrices W_T are assumed to be upper bounded and to be lower bounded away from zero. This assumption is common in the literature, especially through a maintained assumption about the variance matrix of the moment conditions when the weighting matrix is chosen to be a consistent estimator of its inverse. More generally speaking, no moment condition (or linear combination of moment conditions) is endowed with a weight asymptotically negligible with respect to the weight of the other moment conditions. It is important to realize that without such an assumption, the heterogeneity of identification strengths of the moment conditions would not be well defined. To see that, let us define the rescaled moment conditions:

$$\bar{\zeta}_T(\theta) = \sqrt{T}\Lambda_T^{-1}\bar{\psi}_T(\theta). \tag{2.4}$$

For sake of expositional simplicity, let us consider the case where the scaling matrix Λ_T is diagonal.

Example 1 (ctnd): GMM with Many Moment Conditions and Diagonal Λ_T

It would then be natural to consider GMM estimators of θ based on the moment conditions $\bar{\zeta}_T(\theta)$, for instance the simplest one computed with an identity weighting matrix:

$$\breve{\theta}_T = \arg\min_{\theta} \bar{\zeta}_T(\theta)' \bar{\zeta}_T(\theta) = \arg\min_{\theta} \sum_{j=1}^{J_T} \frac{\bar{\psi}_{j,T}^2(\theta)}{\lambda_{j,T}^2}$$

Obviously, when there is no heterogeneity of identification strengths ($\lambda_{j,T} = \lambda_T, \forall j = 1, ..., J_T$), $\breve{\theta}_T$ is a GMM estimator in the sense of definition 1 with an identity weighting matrix W_T. In contrast, in case of heterogeneity, for instance:

$$\lambda_{j,T} = T^{\gamma_j},\ 0 < \gamma_j \leq \frac{1}{2}$$

we can see $\breve{\theta}_T$ as:

$$\breve{\theta}_T = \arg\min_{\theta} \bar{\psi}_T(\theta)' W_T \bar{\psi}_T(\theta),$$

where $W_T = \Lambda_T^{-1}$ is the diagonal matrix with diagonal coefficients $\left(1/T^{2\gamma_j}\right),\ j = 1, ..., J_T$. Obviously, when the exponents $\gamma_j,\ j = 1, ..., J_T$ are not all equal, there does not exist a real sequence ω_T such that the sequence of matrices $W_T^* = \omega_T W_T$ has eigenvalues that are upper bounded and lower bounded away from zero. In other words, the estimator $\breve{\theta}_T$ is not a GMM estimator based on the moment functions $\bar{\psi}_T(\theta)$ in the sense of definition 1.

More generally, it is worth noting that $\breve{\theta}_T$, albeit possibly not a GMM estimator based on the moment functions $\bar{\psi}_T(\theta)$ (in the sense of definition 1), may be a meaningful consistent GMM estimator of θ, but (in the sense of definition 1) based on the moment conditions $\bar{\zeta}_T(\theta)$. Actually, we have by definition:

$$E\left[\bar{\zeta}_T(\theta)\right] = \rho_T(\theta)$$

and our main identification assumption will be relevant for both GMM estimators based on $\bar{\psi}_T(\theta)$ or on $\bar{\zeta}_T(\theta)$. Following (2.3), we will assume:

Assumption 1: We have, with uniform convergence on Θ:

$$\lim_{T \longrightarrow \infty} \frac{\|\rho_T(\theta)\|^2}{J_T} = \delta(\theta)$$

with:

(i) (global identification):

$$\delta(\theta) = 0 \Longleftrightarrow \theta = \theta^0,$$

(ii) (well-separated zero):

$$\forall \varepsilon > 0, \inf_{\|\theta - \theta^0\| > \varepsilon} \delta(\theta) > 0,$$

(iii) (local misspecification):

$$\frac{\left\|\Lambda_T \rho_T\left(\theta^0\right)\right\|}{\sqrt{J_T}} = O(1).$$

As already mentioned, Assumption 1 involves the uniform convergence in (2.3) that coincides with

Assumption 2(i) in HP. The second part of Assumption 1 is a condition of asymptotic identification as Assumption 3 in HP. However, we are slightly less general by setting the focus on identification through the main signal (the moment conditions) while overlooking the additional possibility of identification through "signal variability" as considered in HP. Strictly speaking, our identification condition corresponds to Assumption 3 in HP with $\alpha = 0$. However, as announced in the former subsection, HP's coefficient α is for us a function $\alpha(\theta)$ of the unknown parameters and we will not need to assume it uniformly equal to zero for all θ (see the condition (2.10) in theorem 1 below).

Remark: Assumption 1(iii) points out that we may have $\rho_T\left(\theta^0\right) \neq 0$, insofar as:

$$\left\|\Lambda_T \rho_T\left(\theta^0\right)\right\| = O\left(\sqrt{J_T}\right).$$

In particular, in the case of a diagonal matrix Λ_T:

$$\frac{1}{J_T}\sum_{j=1}^{J_T} \lambda_{j,T}^2 \rho_{j,T}^2\left(\theta^0\right) = O(1).$$

In other words, in average, local misspecification of the moment conditions means that:

$$\rho_{j,T}\left(\theta^0\right) = O\left(\frac{1}{\lambda_{j,T}}\right).$$

Not surprisingly, the local misspecification may be more severe than the standard definition of local misspecification in case of root-T asymptotics (see for instance section 5.1.3 in Hall (2005)). The side effect of nearly weak identification ($\lambda_{j,T} \ll \sqrt{T}$) is that estimation will not be contaminated by a bias of order $1/\sqrt{T}$.

To some extent, Assumption 1 makes even more natural a GMM approach based (in the sense of definition 1) on the moment conditions $\bar{\zeta}_T(\theta)$ since the population identifying function is:

$$\delta(\theta) = \lim_{T \longrightarrow \infty} \frac{\left\|E\left[\bar{\zeta}_T(\theta)\right]\right\|^2}{J_T} \tag{2.5}$$

From (2.5), one may have the spurious feeling that identification is not impaired by any degree of weakness. Example 2.2 below shows that on the one hand, the choice between GMM based on $\bar{\psi}_T(\theta)$ or GMM based on $\bar{\zeta}_T(\theta)$ (always in the sense of definition 1) is actually feasible, but there may be good reasons to prefer GMM based on $\bar{\psi}_T(\theta)$. More precisely, we will argue that the empirical process of the relevant moment functions $\bar{\psi}_T(\theta)$ is properly bounded by a tightness condition.

Assumption 2: If $\Phi_T(\theta)$ stands for the empirical process:

$$\Phi_T(\theta) = \sqrt{T}\left[\bar{\psi}_T(\theta) - \frac{\Lambda_T}{\sqrt{T}}\rho_T(\theta)\right],$$

we have:

$$\sup_{\theta\in\Theta} \frac{\|\Phi_T(\theta)\|^2}{J_T} = O_P(1).$$

Assumption 2 means that the centered and rescaled moment functions of interest $\bar{\psi}_T(\theta)$ are uniformly tight, and it is satisfied in particular if $\bar{\psi}_T(\theta)$ follows a functional central limit theorem. This assumption, first introduced in the weak identification literature by Stock and Wright (2000), has been used by HP and AR12 to prove consistency of the GMM estimators of interest. As emphasized by HP, the consistency proof is based in the fact that assumption 2 implies that $\left[\bar{\psi}_T(\theta)'W_T\bar{\psi}_T(\theta)/J_T\right]$ is also tight, so that we can prove the following lemma.

Lemma 1: Under Assumptions 1 and 2, a GMM estimator $\hat{\theta}_T$ of θ based on the moment functions $\bar{\psi}_T(\theta)$ is such that:

$$\left\|\Lambda_T\rho_T\left(\hat{\theta}_T\right)\right\| = O_P\left(\sqrt{J_T}\right).$$

Example 2: Local GMM via Kernel Smoothing

Example 2.1: Local GMM

Efficient estimation of model with conditional moment restrictions containing unknown functions has been studied since Ai and Chen (2003). However, Lewbel (2007) pointed out that this estimation problem becomes much simpler if one assumes that the unknown functions are only functions of the conditioning variables of the conditional moment restrictions. In other words, we are interested in conditional moment restrictions written as

$$E[g(Y,\theta(z))\,|Z=z] = 0. \tag{2.6}$$

In this example 2.1, we will maintain as Lewbel (2007) the assumption that we can observe an independently, identically distributed random sample of observations $Y_t, Z_t, t = 1, ..., T$ of the random vectors Y, Z. Extensions to stationary time series will be considered in example 2.2 below. The key idea is then to realize that we may have at our disposal a consistent estimator of the conditional moment in (2.6). Let us consider a Nadaraya-Watson estimator

$$\hat{E}_T[g(Y,\theta)\,|Z=z] = \left[\sum_{t=1}^{T} K\left(\frac{Z_t - z}{h_T}\right)\right]^{-1} \sum_{t=1}^{T} g(Y_t,\theta)\,K\left(\frac{Z_t - z}{h_T}\right), \tag{2.7}$$

where K is a nonnegative symmetric kernel function satisfying

$$\int K(u)du = 1, \quad \int \|K(u)\|^r du < \infty, r > 2.$$

Asymptotic theory will take some regularity conditions stated as follows in Lewbel (2007).

(i) The d vector Z is continuously distributed with density function $f(z)$.

(ii) For some z in the interior of the support of Z having $f(z) > 0$, $g(Y, \theta)$ is twice continuously differentiable in the vector θ for all θ in some compact subset $\Theta(z)$ of $\Theta \subset \mathbb{R}^p$.

In addition, Lewbel (2007) maintains the following identification assumptions.

Assumption (Identification Assumptions):

(i) For some z in the interior of the support of Z having $f(z) > 0$, there exists a unique $\theta^0(z)$ in $\Theta(z)$ such that

$$E[g\left(Y, \theta^0(z)\right) | Z = z] = 0.$$

(ii) The matrix

$$E\left[\frac{\partial g\left(Y, \theta^0(z)\right)}{\partial \theta'} | Z = z\right]$$

is of rank p.

We can then define a local GMM estimator of $\theta^0(z)$ as the minimizer of the sample objective function

$$Q_T(\theta) = \left[\hat{E}_T[g(Y, \theta) | Z = z]\right]' W_T \left[\hat{E}_T[g(Y, \theta) | Z = z]\right]$$

for some sequence of random positive definite matrices W_T converging toward a positive definite matrix W.

It is worth mentioning that the exposition in Lewbel (2007) is slightly different since he rightly considers that since the solution of the above minimization is not modified by a positive scaling factor, the scaling of the kernel weights in (2.7) to make them summing to 1 is immaterial. In other words, he does not really use the Nadaraya-Watson estimator but only the weighted sum

$$\sum_{t=1}^{T} g(Y_t, \theta) K\left(\frac{Z_t - z}{h_T}\right).$$

This difference explains that we will present our assumptions and results in a slightly different way, but, once more, the difference is immaterial. It is more convenient for us to use explicitly a consistent estimator of the conditional expectation in order to figure out the relevant generalization in example 2.2 when we also incorporate some standard moment restrictions. Obviously, the tightness

assumption 2 takes:

$$\bar{\psi}_T(\theta) = \sqrt{h_T^d}\hat{E}_T[g(Y,\theta)\,|Z=z]$$

Under standard regularity conditions, for a fixed J_T, if $Th_T^d \to \infty$ and $Th_T^{d+4} \to 0$:

$$\sqrt{Th_T^d}\left[\hat{E}_T[g(Y,\theta)\,|Z=z] - E[g(Y,\theta)\,|Z=z]\right] \longrightarrow^d \aleph\,(0, V(z))$$

for some variance matrix $V(z)$ (see standard kernel regression limiting distribution theory). In comparison with assumption 2, we see that if J_T stands for the dimension of the moment function g:

$$\Lambda_T = \lambda_T Id_{J_T}, \lambda_T = \sqrt{Th_T^d}$$

so that:

$$c_T = J_T h_T^d.$$

Hence, Lewbel (2007)'s main result (Theorem 1) which gives consistency and asymptotic normality of a GMM estimator of θ with a rate of convergence $\sqrt{Th_T^d}$ can be reinterpreted as a particular case (in the case of a fixed number J_T of moment conditions) of the general result of HP (see corollary 13) which proves asymptotic normality at rate $\sqrt{Tc_T}$. Note that, for fixed J_T, our condition of local misspecification (Assumption 1(iii)), states that:

$$\lambda_T \rho_T\left(\theta^0\right) = \sqrt{Th_T^d}\rho_T\left(\theta^0\right) = O(1)$$

that is nothing but the common assumption that the kernel smoothing bias, rescaled by $\sqrt{Th_T^d}$, does not blow up. By assuming $Th_T^{d+4} \to 0$, we make bias shrink faster than variance.

Example 2.2: Kernel Moment Estimator

Following Gagliardini et al. (2011), we are interested in conditional moment restrictions that generalize (2.6) because they can be written as

$$\begin{aligned} E[g\,(Y,\theta)\,|Z=z] &= 0,\ a.s.\ \forall z \in S_Z \\ E[\tilde{g}\,(Y,\theta)\,|Z=z_0] &= 0, \end{aligned} \tag{2.8}$$

where S_Z stands for the support of the random vector Z and $a.s.$ means "almost surely" for the probability distribution of Z. z_0 stands for a specific value of Z that belongs to the set of values observed in our sample. Gagliardini et al. (2011) actually generalize also the setting of local GMM by considering observations $(Y_t, Z_t), t = 0, 1, \cdots, T$ on a stationary time series. Then z_0 is the value observed for Z_0. We maintain the following identification assumptions.

Assumption (Global Identification): There exists a unique true value θ^0 of the unknown parameters $\theta \in \Theta$ such that, on the one hand

$$E[g\left(Y, \theta^0\right) | Z = z] = 0, \ a.s. \ \forall z \in S_Z,$$

and on the other hand

$$E[\tilde{g}\left(Y, \theta^0\right) | Z = z_0] = 0.$$

Assumption (Local Identification): There exists no non-zero vector $\alpha \in \mathbb{R}^p$ such that

$$\begin{aligned} E\left[\frac{\partial g\left(Y, \theta^0\right)}{\partial \theta'} | Z = z\right].\alpha &= 0, a.s. \forall z \in S_Z \\ E\left[\frac{\partial \tilde{g}\left(Y, \theta^0\right)}{\partial \theta'} | Z = z_0\right].\alpha &= 0. \end{aligned}$$

These identification assumptions strengthen to some extent the identification assumptions of the local GMM estimator seen as assumptions about the complete set of moment conditions

$$g_2(Y, \theta) = \left[\ g(Y, \theta)' \ \tilde{g}\left(Y, \theta\right)' \ \right]'.$$

While the assumption in example 2.1 would have led to identify a different parameter value $\theta(z)$ for each possible conditioning value z, the above global and local identification assumptions maintain that $\theta(z)$ and $\theta^0 = \theta(z_0)$ have some features in common such that θ^0 is solution of the moment conditions associated to $g(Y, \theta)$ *uniformly* over the possible values of the conditioning variables Z. Note, however, that the above assumptions do not strictly speaking imply that $\theta^0 = \theta(z_0)$ is identified by considering only conditional moment restrictions given $Z = z_0$.

However, one expects that a local GMM estimator of $\theta^0 = \theta(z_0)$, obtained as minimizer of the sample objective function

$$Q_{2,T}(\theta) = \left[\hat{E}_T[g_2\left(Y, \theta\right) | Z = z_0]\right]' W_T \left[\hat{E}_T[g_2\left(Y, \theta\right) | Z = z_0]\right] \tag{2.9}$$

may be consistent but likely less accurate that an estimator that would take into account the fact that some conditional moment restrictions are valid uniformly over the possible values of the conditioning variables Z. As usual, these uniform conditional moment restrictions can be handled for the purpose of GMM estimation through the choice of a convenient matrix $H(Z)$ of instruments. More precisely, introducing the moment functions

$$g_1(Z, Y, \theta) = H(Z)g(Y, \theta),$$

we will say that these instruments are *admissible* if and only if on the one hand, we have the two set of conditions

$$E\left[g_1(Z,Y,\theta)\right]=0,\ E[g_2\left(Y,\theta\right)|Z=z_0]=0$$

if and only if $\theta=\theta^0$, and on the other hand, there exists no non-zero vector $\alpha\in\mathbb{R}^p$ such that

$$E\left[\frac{\partial g_1\left(Z,Y,\theta^0\right)}{\partial\theta'}\right].\alpha=0,\ E\left[\frac{\partial g_2\left(Y,\theta^0\right)}{\partial\theta'}|Z=z_0\right].\alpha=0.$$

The global and local identification assumptions together imply the existence of admissible instruments. With such instruments in hands, we should intuitively improve the asymptotic accuracy of our estimator by replacing local GMM based on $g_2(Y,\theta)$ by a GMM estimator based on the moment functions $\bar{\psi}_T(\theta)$ in the sense of definition 1 with:

$$\begin{aligned}\bar{\psi}_T(\theta) &= \left[\bar{\psi}_{1,T}(\theta)',\bar{\psi}_{2,T}(\theta)'\right]'\\ \bar{\psi}_{1,T}(\theta) &= \hat{E}_T[g_1(Z,Y,\theta)]=\frac{1}{T}\sum_{t=1}^{T}g_1(Z_t,Y_t,\theta)\\ \bar{\psi}_{2,T}(\theta) &= \sqrt{h_T^d}\hat{E}_T[g_2(Y,\theta)|Z=z_0].\end{aligned}$$

Under standard regularity conditions, if $Th_T^d\to\infty$ and $Th_T^{d+4}\to 0$,if $J_{1,T}$ (resp. $J_{2,T}$) stands for the dimension of g_1 (resp. g_2) we see that:

$$\Lambda_T=\begin{bmatrix}Id_{J_{1,T}} & 0\\ 0 & \lambda_T Id_{J_{2,T}}\end{bmatrix},\ \lambda_T=\sqrt{Th_T^d}.$$

Of course, one may even consider more general diagonal matrices Λ_T, by choosing different bandwidth parameters for the different components of g_2. Note that, possibly by contrast with the weak identification example, kernel moment estimation is a case where the different degrees of weakness are actually known by the researcher (corresponding to the choice of bandwidth parameters). Thus, the researcher may also consider the rescaled moments as in (2.4):

$$\begin{aligned}\bar{\zeta}_T\left(\theta\right) &= \left[\bar{\zeta}_{1,T}\left(\theta\right)',\bar{\zeta}_{2,T}\left(\theta\right)'\right]'\\ \bar{\zeta}_{1,T}\left(\theta\right) &= \hat{E}_T[g_1(Z,Y,\theta)]\\ \bar{\zeta}_{2,T}\left(\theta\right) &= \hat{E}_T[g_2(Y,\theta)|Z=z_0].\end{aligned}$$

A GMM estimator based on the moment functions $\bar{\zeta}_T\left(\theta\right)$ in the sense of definition 1 would be consistent by virtue of standard arguments, at least for a fixed J_T. However, a GMM based

on $\bar{\zeta}_T(\theta)$ should be less efficient than a GMM based on $\bar{\psi}_T(\theta)$ for the following reason. When GMM is based on $\bar{\zeta}_T(\theta)$ in the sense of definition 1, it means that moment conditions provided by $\hat{E}_T[g_2(Y,\theta)\,|Z=z_0]$ are endowed with a weight that is not negligible in front of weights assigned to $\hat{E}_T[g_1(Z,Y,\theta)]$. This cannot be optimal for asymptotic variance of GMM estimators since the variance of $\hat{E}_T[g_2(Y,\theta)\,|Z=z_0]$ is asymptotically infinitely larger than the variance of $\hat{E}_T[g_1(Z,Y,\theta)]$.

2.3 Consistent GMM estimation

It has been well-known since Chao and Swanson (2005) that perverse correlation due to endogeneity may not disappear asymptotically when the number of moment conditions goes to infinity so that 2SLS is biased even asymptotically, except if we assume (see their theorem 2.4) that the concentration parameter $\bar{\lambda}_T^2(\theta)J_T$ goes to infinity faster than the number J_T of moment conditions, that is $\bar{\lambda}_T^2(\theta)$ goes to infinity (it does not depend on θ in the linear IV model of Chao and Swanson (2005), see Appendix B). As noted by Newey and Windmeijer (2009) (see page 694), this problem also kicks in for (nonlinear) two-step GMM due to perverse correlation between Jacobian matrix and moment conditions. Alternative approaches to erase this bias (LIML, Continuously Updated GMM, Empirical Likelihood) are beyond the scope of this paper. We are following HP by setting the focus on GMM with a given weighting matrix in the sense of definition 1.

The condition that $\bar{\lambda}_T^2(\theta)\to\infty$, maintained in Chao and Swanson (2005) for the consistency of 2SLS, is also maintained to a large extent by HP by assuming that $\alpha_T(\theta)\to 0$, ($\alpha_T(\theta)=1/\bar{\lambda}_T^2(\theta)$) even though they do not take into account the dependence of $\alpha_T(\theta)$ on θ since they do not address the case of heterogeneity of identification strengths. We acknowledge that HP also discuss identification for a non-zero finite limit of $\alpha_T(\theta)$, but this case is beyond our scope since it is based on identification through signal variability. As a matter of fact, when it comes to get a GMM estimator that is asymptotically normal with a "textbook asymptotic variance matrix", HP are actually led to assume (see their corollary 13) that $\alpha_T(\theta)\to 0$.

The good news is that in our context of a generalized concentration parameter that depends on θ due to heterogeneity of identification strengths, only the value of this concentration parameter, computed at the GMM estimator, matters for consistency of this estimator. We define more generally:

$$\bar{\lambda}_T^2(\theta)=\frac{\rho_T(\theta)'\,\Lambda_T\rho_T(\theta)}{\rho_T(\theta)'\,\rho_T(\theta)}.$$

As extensively discussed in subsection 2.1, $\bar{\lambda}_T^2(\theta)$ can be interpreted, at least in the case of a diagonal matrix Λ_T, as an average value of identification strengths of the various moment equations, with weights that are dependent on θ. This concept allows us to prove a consistency theorem for a GMM estimator that complements theorem 5 of HP for the case of heterogeneous identification strengths.

Our proof is completely different even though maintained assumptions are very similar.

Theorem 1: Under Assumptions 1 and 2, if $\hat{\theta}_T$ is a GMM estimator of θ based on the moment functions $\bar{\psi}_T(\theta)$ that is such that:

$$\lim_{T \longrightarrow \infty} \bar{\lambda}_T^2(\hat{\theta}_T) = +\infty, \tag{2.10}$$

then, $\hat{\theta}_T$ is a weakly consistent estimator of θ:

$$P \lim_{T \longrightarrow \infty} \hat{\theta}_T = \theta^0.$$

Remark: When used to prove consistency of the GMM estimator $\hat{\theta}_T$, the condition (2.10) is arguably a high level assumption since it is already about the asymptotic behavior of our GMM estimator of interest $\hat{\theta}_T$. However, this assumption about the average identification strength, is far from imposing that all the moment conditions are not genuinely weak. Imagine for instance that the J_T moment conditions can be partitioned in two subsets of $J_{1,T}$ and $J_{2,T}$ conditions respectively such that the matrix is block diagonal:

$$\Lambda_T = \begin{bmatrix} \Lambda_{1,T} & 0 \\ 0 & \Lambda_{2,T} \end{bmatrix}.$$

Then with obvious notations:

$$\rho_T(\theta)' \Lambda_T \rho_T(\theta) = \rho_{1,T}(\theta)' \Lambda_{1,T} \rho_{1,T}(\theta) + \rho_{2,T}(\theta)' \Lambda_{2,T} \rho_{2,T}(\theta).$$

It is then clear that if the first subset of $J_{1,T}$ moment conditions fulfills the condition (2.10), we get the required condition (2.10) for the full set of J_T conditions as long as:

$$\frac{\left\| \rho_{2,T}\left(\hat{\theta}_T\right) \right\|^2}{\left\| \rho_{1,T}\left(\hat{\theta}_T\right) \right\|^2} = O_P(1). \tag{2.11}$$

Moreover, while conditions (2.10) and (2.11) have been written at the GMM estimator, it is of course sufficient to maintain them uniformly over the parameter set Θ.

2.4 Another example: Near Multicollinearity

It is worth realizing that our general framework not only nests the standard GMM framework with nearly weak identification and the kernel moment estimation but also some issues that are not in general seen as pertaining to the realm of weak identification. We have in mind the case

of partially linear models (PLM) with many regressors where the non-standard asymptotic theory may kick in not only due to the large number of regressors, but also due to the fact that the steady increase of the number of regressors may create some near multicollinearity issue.

We depart from the models with many instruments considered in CS and the standard PLM with many regressors studied in particular by Anatolyev (2012) and Cattaneo et al. (2018), by considering cases where the weakness of moment conditions arises from the near-multicollinearity of explanatory variables rather than the near-irrelevance of some instruments.

We are interested in a partially linear equation

$$y_t = x_t'\theta^0 + g^0(w_t) + u_t,\ x_t \in \mathbb{R}^p,$$

where $g^0(w_t)$ is an unknown function that can be approximated by a series expansion on a basis of functions: $p^1(w_t)$, $p^2(w_t)$, ... More precisely, we introduce the K-dimensional function

$$p_K(w) = \left[p^1(w), p^2(w), ..., p^K(w)\right]'$$

and for as sequence $K = K_T$ going to infinity with T, a sequence of approximations $p'_{K_T}(w)\beta_T$ of the function $g^0(w)$. We allow the variables x_t and w_t to be endogenous.

Let $z_{t,T}$ denote the J_T-dimensional vector of valid instruments where J_T grows to infinity with the sample size. Similarly with the standard framework of linear model with endogeneity, we write a reduced form regression equation

$$\left[a'_{t,T}, p'_{K_T}(w_t)\right]' = \Pi'_T z_{t,T} + \epsilon_{t,T}, \tag{2.12}$$

where Π_T is a $J_T \times (K_T + p)$ matrix of coefficients with $J_T \geq K_T + p$ and $a_{t,T}$ is a p-dimensional residuals defined from

$$\begin{aligned} x_t &= B_T p_{K_T}(w_t) + v_{t,T}, \\ v_{i,t,T} &= \frac{a_{i,t,T}}{\delta_{i,T}}, \quad \inf_T \min_{1\leq t\leq T} Var(a_{i,t,T}) > 0, \quad \forall i = 1, 2, ..., p. \end{aligned}$$

The term $B_T p_{K_T}(w_t)$ represents the series approximation of the unknown function $E[x_t|w_t]$. When $\delta_{i,T}$ is a finite nonzero constant and Π_T shrinks toward to a zero matrix, this example is a standard PLM with endogeneity with many weak instruments.

However, as stated above, we do not consider the cases with shrinking Π_T. The key idea of this example is that the weak identification of moments is due to the near-multicollinearity of explanatory variables, x_t and $p_{K_T}(w_t)$. This case may arise when a researcher employs a number of series of functions to obtain consistent parameter estimation but such approximating functions also span

the explanatory variables of interest. This near-multicollinearity is accommodated by allowing $\delta_{i,T}$ to diverge with the sample size. Note that this is in contrast with Cattaneo et al. (2018) that maintains that the eigenvalues of $E\left[v_{t,T}v'_{t,T}|w_t\right]$ matrix are bounded away from zero. The following assumption provides the conditions for the rate at which $\delta_{i,T}$ diverges and standard conditions imposed on PLM.

Assumption: (i) $\underset{1\leq i\leq p}{max}\ \delta_{i,T} = o(\sqrt{T})$, (ii) $\inf_T \min_{1\leq i\leq p} \delta_{i,T} > 0$, (ii) The data $(y_t, x_t, w_t), t = 1, 2, ...$ are independent identically distributed (i.i.d.), (iii) There exists $C < \infty$ such that for all $T = 1, 2, ..$ and all $t = 1, 2, ..., T$, we have $E[u_t^4] \leq C$ and $E[\|v_{t,T}\|^4] \leq C$.

The J_T dimensional vector of the instruments $z_{t,T}$, will be considered as valid insofar as

$$E\left[z_{t,T}\left(y_t - x'_t\theta^0 - g^0(w_t)\right)\right] = 0, \tag{2.13}$$

similarly to the standard framework of the PLM with endogeneity. While the moment conditions in (2.13) contain the unknown function $g^0(w_t)$, we will rather consider, in line with Ai and Chen (2003), the approximation of moment conditions provided by a series expansion

$$E\left[z_{t,T}\left(y_t - x'_t\theta^0 - p'_{K_T}(w_t)\beta_T\right)\right] = E\left[z_{t,T}\left(g^0(w_t) - p'_{K_T}(w_t)\beta_T\right)\right] \approx 0.$$

Note that these approximated moments may allow to consistently estimate unknown parameters (θ, β_T) if the number J_T of moment conditions is at least as large as the number $(K_T + p)$ of parameters to estimate. Hence both K_T and J_T must go to infinity with T. Then, the approximation error is assessed directly at the level of the bias for IV estimation. The following assumption is exactly what we need to mimic the order of the bias in our assumption 1(iii).

Assumption (Approximation Error): There exists a sequence β_T^0 such that

$$E\left[z_{t,T}\left(y_t - x'_t\theta^0 - p'_{K_T}(w_t)\beta_T^0\right)\right] = O\left(\frac{\sqrt{J_T}}{\sqrt{T}}\right).$$

In order to handle these moment conditions within our general framework, we proceed as follows. First, we set the focus on the case of two stage least squares, which, with GMM notations, means that we consider the vector of moment conditions

$$\psi^*_{t,T}(\theta, \beta_T) = z_{t,T}\left(y_t - x'_t\theta^0 - p'_{K_T}(w_t)\beta_T\right)$$

and minimize with respect to (θ, β_T) the quadratic function

$$\bar{\psi}_T^*(\theta,\beta_T)'(Z_T'Z_T)^{-1}\bar{\psi}_T^*(\theta,\beta_T), \quad \text{with} \quad \bar{\psi}_T^*(\theta,\beta_T) = \frac{1}{T}\sum_{t=1}^{T}\psi_{t,T}^*(\theta,\beta_T). \tag{2.14}$$

Note that, since it ultimately amounts to OLS on the projection of the x and w variables on the columns of the matrix Z, the Frisch-Waugh theorem allows us to compute in closed form the GMM estimator of θ. Assuming that θ is our parameter of interest, we are back to an inference problem similar to our general setup even though it goes through a GMM step with an infinite number of parameters. As a result, and for sake of expositional simplicity, all the discussions related to this example will be focused on the GMM problem defined by (2.14) even though we are ultimately interested on a finite dimensional parameter θ, which is conformable to our general framework.

Comparison with our general framework

From the moment conditions given above, we can construct the following moment function

$$\begin{aligned}\psi_{t,T}^*(\theta,\beta_T) &= z_{t,T}\left(y_t - x_t'\theta - p_{K_T}'(w_t)\beta_T\right) = z_{t,T}\left[y_t - v_{t,T}'\theta - p_{K_T}'(w_t)\eta_T\right] = \psi_{t,T}(\mu_T)\\ \eta_T &= \beta_T + B_T'\theta,\ \mu_T = (\theta', \eta_T')',\end{aligned}$$

and write

$$\begin{aligned}E\left[\bar{\psi}_T(\mu_T)\right] = E\left[\frac{1}{T}\sum_{t=1}^{T} z_{t,T}[\frac{a_{1,t,T}}{\delta_{1,T}},...,\frac{a_{p,t,T}}{\delta_{p,T}}]\right]\left(\theta^0-\theta\right) + E\left[\frac{1}{T}\sum_{t=1}^{T} z_{t,T}p_{K_T}'(w_t)\right]\left(\eta_T^0-\eta_T\right)\\ + E\left[\frac{1}{T}\sum_{t=1}^{T} z_{t,T}\left\{g^0(w_t) - p_{K_T}(w_t)'\beta_T^0\right\}\right].\end{aligned}$$

From the reduced form equation given in (2.12), we can deduce that

$$E\left[\bar{\psi}_T(\mu_T)\right] = E\left[z_{t,T}z_{t,T}'\right]\Pi_T\frac{\bar{\Lambda}_T}{\sqrt{T}}\left(\mu_T^0-\mu_T\right) + E\left[z_{t,T}\left\{g^0(w_t) - p_{K_T}(w_t)'\beta_T^0\right\}\right], \tag{2.15}$$

where $\bar{\Lambda}_T$ is a diagonal matrix of size $(p+K_T)$ whose first p diagonal coefficients are $\sqrt{T}/\ \delta_{i,T}, i = 1,..,p$, while the last K_T coefficients are all equal to $\sqrt{T}$.

In order to see that the moments given in (2.15) can be written as our general form, let $E\left[z_{t,T}z_{t,T}'\right]$ be a diagonal matrix[1]. Also let Λ_T a diagonal matrix of dimension J_T that has the same largest

[1]Note that this assumption can be relaxed to $E\left[z_{t,T}z_{t,T}'\right]$ being a positive definite matrix for all T but we maintain this for expositional simplicity. See the supplemental appendix B with the example of CS for more detail.

and smallest coefficients as $\bar{\Lambda}_T$, similarly to lemma 4.1. of AR12. Then we can write

$$E\left[z_{t,T}z_{t,T}'\right]\Pi_T\frac{\bar{\Lambda}_T}{\sqrt{T}}(\mu_T^0-\mu_T)=E\left[z_{t,T}z_{t,T}'\right]\frac{\Lambda_T}{\sqrt{T}}\sqrt{T}\Lambda_T^{-1}\Pi_T\frac{\bar{\Lambda}_T}{\sqrt{T}}(\mu_T^0-\mu_T)$$
$$=\frac{\Lambda_T}{\sqrt{T}}E\left[z_{t,T}z_{t,T}'\right]\Lambda_T^{-1}\Pi_T\bar{\Lambda}_T(\mu_T^0-\mu_T),$$

so that $E\left[\bar{\psi}_T(\mu_T)\right]=\Lambda_T\rho_T(\mu_T)/\sqrt{T}$ with

$$\rho_T(\mu_T)=E\left[z_{t,T}z_{t,T}'\right]\Lambda_T^{-1}\Pi_T\bar{\Lambda}_T(\mu_T^0-\mu_T)+\sqrt{T}\Lambda_T^{-1}E\left[z_{t,T}\left\{g^0(w_t)-p_{K_T}(w_t)'\beta_T^0\right\}\right].$$

Note that the smallest diagonal coefficient of $\bar{\Lambda}_T$ goes to infinity with T by the first assumption made in this example. Hence, the condition (2.10) in theorem 1 is fulfilled.

Note also that

$$\Lambda_T\rho_T(\mu_T^0)=\sqrt{T}E\left[z_{t,T}\left\{g^0(w_t)-p_{K_T}(w_t)'\beta_T^0\right\}\right]$$

so that our assumption 1(iii) is implied by the above approximation error assumption. We also assume, by extending to our setting of a number J_T of moments going to infinity with the sample size the result of theorem 2.3. page 40 of Antoine and Renault (2011), that

$$\lim_{T\to\infty}\bar{\Lambda}_T\Pi_T'\Lambda_T^{-1}E\left[z_{t,T}z_{t,T}'\right]\Lambda_T^{-1}\Pi_T\bar{\Lambda}_T/J_T$$

exists and is positive definite. Hence, our assumption 1(i) and (ii) also hold.

See that the sample counterpart of the moments is

$$\bar{\psi}_T(\mu_T)=\frac{1}{T}\sum_{t=1}^T z_{t,T}z_{t,T}'\Pi\frac{\bar{\Lambda}_T}{\sqrt{T}}(\mu_T^0-\mu_T)+\frac{1}{T}\sum_{t=1}^T z_{t,T}\varepsilon_{t,T}'\frac{\bar{\Lambda}_T}{\sqrt{T}}(\mu_T^0-\mu_T)$$
$$+\frac{1}{T}\sum_{t=1}^T z_{t,T}u_{t,T}+\frac{1}{T}\sum_{t=1}^T z_{t,T}\left[g^0(w_t)-p_{K_T}'(w_t)\beta_T^0\right].$$

We assume the following conditions to hold.

Assumption (Uniform Boundedness): (i) $\left\|T^{-1/2}\sum_{t=1}^T z_{t,T}\varepsilon_{t,T}'\right\|/\sqrt{J_T}=O_p(1)$, (ii) $\left\|T^{-1/2}\sum_{t=1}^T z_{t,T}u_{t,T}\right\|/\sqrt{J_T}=O_p(1)$, (iii) $\left\|\sqrt{T}\left[E\left[z_{t,T}z_{t,T}'\right]-\frac{1}{T}\sum_{t=1}^T z_{t,T}z_{t,T}'\right]\Pi_T\right\|/\sqrt{J_T}=O_p(1)$,
(iv) $\left\|T^{-1/2}\sum_{t=1}^T\left[z_{t,T}\left\{g^0(w_t)-p_{K_T}(w_t)'\beta_T^0\right\}-E\left[z_{t,T}\left\{g^0(w_t)-p_{K_T}(w_t)'\beta_T^0\right\}\right]\right]\right\|/\sqrt{J_T}=O_p(1)$.

The above assumption implies our assumption 2. Note that since the moments of interest are linear with respect to the parameters, there is no need to consider conditions with uniformity over Θ.

Since all conditions in theorem 1 are verified to hold (for an increasing number of parameters), we can conclude that $\hat{\theta}_T$, the GMM estimator of θ which is the first p components of μ_T and the parameter of interest, is weakly consistent.

3 Asymptotic Normality of GMM Estimators

The goal of this section is to encompass in our unified setting the asymptotic distributional theory of GMM based on many weak moments (like HP) but with heterogeneity of instrument strength (AR12). For sake of expositional simplicity, we maintain the simplifying assumption of example 1, namely a diagonal Λ_T matrix.

3.1 Minimum rate of convergence

In order to unify the contributions of HP and AR12, two messages must be taken into account.

On the one hand, with a finite number of moments, AR12 shows that the diagonal coefficients $\lambda_{j,T}$ of the matrix Λ_T determine the rates of convergence of GMM estimators of different linear combinations of the unknown parameters θ. If

$$\lambda_{i+1,T} = o\left(\lambda_{i,T}\right), i = 1, 2, ..., l-1, \tag{3.1}$$

theorem 4.3 of AR12 shows that after a convenient reparameterization $\eta = R\theta$, $\lambda_{i,T}, i = 1, 2, ...l$ is the rate of convergence of GMM estimators of some components of η. If for instance $\lambda_{1,T} = \sqrt{T}$, some components of η are strongly identified while nearly weak/strong identification of some other components may be at stake with smaller rates of convergence $\lambda_{2,T}, \lambda_{3,T}, \cdots$.

On the other hand, with an infinite number of moments, but only one degree of (nearly) weakness of the moment conditions, HP (see their theorem 12) shows that the rate of convergence of a GMM estimator is $\sqrt{Tc_T}$ with their notations, that is $\lambda_T\sqrt{J_T}$ with our notations (see example 1).

Therefore, the convenient generalization of the ranking (3.1) of identification strengths is the following ranking, where weakness of some subsets of moment conditions may be compensated by the numbers of moments that are involved

$$\lambda_{i+1,T}\sqrt{J_{i+1,T}} = o\left(\lambda_{i,T}\sqrt{J_{i,T}}\right), i = 1, 2, ..., l-1, \tag{3.2}$$

where

$$J_T = \sum_{i=1}^{l} J_{i,T}$$

$$\Lambda_T = \begin{bmatrix} \lambda_{1,T} Id_{J_{1,T}} & 0 & 0 & 0 \\ 0 & \lambda_{2,T} Id_{J_{2,T}} & 0 & 0 \\ ... & ... & ... & ... \\ 0 & 0 & 0 & \lambda_{l,T} Id_{J_{l,T}} \end{bmatrix} \qquad (*)$$

$$\rho_T(\theta) = [\rho_{iT}(\theta)]_{1 \leq i \leq l}, \ \bar{\psi}_T(\theta) = \left[\bar{\psi}_{iT}(\theta)\right]_{1 \leq i \leq l},$$

where Id_H stands for the identity matrix of dimension H. In the standard setup of linear IV (CS), this ranking is defined by the ranking of concentration parameters $\lambda_{i,T}^2 J_{i,T}$, from the most divergent one $\lambda_{1,T}^2 J_{1,T}$ (meaning the maximum identification strength) to the least divergent one $\lambda_{l,T}^2 J_{l,T}$ (meaning the smallest identification strength). As CS (see their assumption 1(c)), the concentration parameter must diverge no faster than $\sqrt{T}$ (our assumption 3(i) below). Moreover, up to a slight change of notation, the $J_T = \sum_{i=1}^{l} J_{i,T}$ moment conditions corresponding to (3.2) do not necessarily encompass all the moment conditions described in section 2, but only the ones that matter for identification (see more precisely section 3.2 below). This is the reason why we maintain (our assumption 3(ii) below) that even in the case of the smallest identification strength (in this subset of $J_T = \sum_{i=1}^{l} J_{i,T}$ moment conditions) the generalized concentration parameter still diverges with the sample size.

Assumption 3: There exists $l \geq 1$ such that for all T, $J_T = \sum_{i=1}^{l} J_{i,T}$, Λ_T can be decomposed as in $(*)$ and (i) $\lambda_{1,T}\sqrt{J_{1,T}} = O\left(\sqrt{T}\right)$, (ii) $\lim_{T\to\infty} \lambda_{l,T} J_{l,T} = +\infty$.

Assumption 4 below corresponds very closely (see comments on example 1 continued in the appendix A for a more thorough comparison) to the assumptions maintained by HP in their proposition 10, when their assumptions, applied for a rate of convergence $\lambda_T\sqrt{J_T}$ ($\sqrt{Tc_T} = r_T$ in their notations) are adapted for the minimum rate of convergence $\lambda_{l,T}\sqrt{J_{l,T}}$. As HP acknowledges, the second part of assumption 4 is a high-level assumption that is hard to verify. We show that this "Lipschitz condition" holds for the linear IV model with many instruments under the framework of CS in Appendix B in the supplemental appendix.

Assumption 4: Let θ_T^* be defined as, for each T,

$$\theta_T^* = \underset{\theta \in \Theta}{argmin} \sum_{i=1}^{l} \frac{1}{J_{i,T}} \rho_{iT}(\theta)' W_{iT} \rho_{iT}(\theta).$$

where θ_T^* is the unique minimizer and, for all $i = 1, \cdots, l$, W_{iT} is a $J_{i,T} \times J_{i,T}$ positive definite matrix with $\|W_{iT}\| = O_p(1)$ and $\|W_{iT}^{-1}\| = O_p(1)$.

Suppose that there exist an $T_0 < \infty$ and neighborhoods of θ_T^* such that for every θ in the neighborhoods and for fixed constants C,

(i)

$$\|\theta - \theta_T^*\|^2 \leq \sum_{i=1}^{l} \frac{C}{J_{i,T}} \left[\rho_{iT}(\theta)' W_{iT} \rho_{iT}(\theta) - \rho_{iT}(\theta_T^*)' W_{iT} \rho_{iT}(\theta_T^*)\right], \ T > T_0.$$

(ii) Also, for all $T > T_0$ and for every sufficiently small $\epsilon > 0$, we have

$$E\left[\sup_{\|\theta - \theta_T^*\| < \epsilon} \sum_{i=1}^{l} \left| \lambda_{l,T} \sqrt{J_{l,T}} X_{iT}(\theta) - \lambda_{l,T} \sqrt{J_{l,T}} X_{iT}(\theta_T^*) \right| \right] \leq C\epsilon,$$

where

$$X_{iT}(\theta) = \frac{T}{\lambda_{i,T}^2 J_{i,T}} \bar{\psi}_{iT}(\theta)' W_{iT} \bar{\psi}_{iT}(\theta) - \frac{1}{J_{i,T}} \rho_{iT}(\theta)' W_{iT} \rho_{iT}(\theta).$$

Note that with the heterogeneous convergence rates, assumption 4(ii) above imposes only $\lambda_{l,T}\sqrt{J_{l,T}} X_{iT}(\cdot)$ to remain tight for all $i = 1, \cdots, l$. On the other hand, the condition for the cases with a single convergence rate is tantamount to requiring each $\lambda_{i,T}\sqrt{J_{i,T}} X_{iT}(\cdot)$ to be tight. The GMM estimator would attain a faster rate of convergence if we were able to impose assumption 4(ii) with the faster rate than $\lambda_{l,T}\sqrt{J_{l,T}}$ as a scale factor. However, as we can see from the example of CS in the supplemental appendix, the conditions about the expectation in assumption 4(ii) would in general not hold since it would diverge to infinity if faster rates of convergences are applied.

Finally, we need to strengthen assumption 1(iii) as follows.

Assumption 5:

$$\left\|\Lambda_T \rho_T(\theta^0)\right\| = O(1).$$

In other words, in contrast with assumption 1(iii), we no longer want to see the norm of the (rescaled) J_T-dimensional vector of bias to blow up when its dimension is going to infinity. This limitation is the price to pay to take advantage of the large number of moment conditions to improve the rate of convergence. We cannot allow the average local bias to grow when rescaled by this rate of convergence. This strengthening of assumption 1 sounds quite reasonable, either because we allow for non-zero bias only in a bounded number of moments (like for instance in the context of kernel smoothing bias) or, as in example 3, we consider that when the number of moment grows (with

T), the bias in each moment tends to decrease (more accurate approximation with longer series expansions in the context of example 3).

With these maintained assumptions we are able to prove theorem 2 below.

Theorem 2 (minimum rate of convergence of a GMM estimator): Under the conditions in theorem 1 and assumptions 3 to 5,

$$\left\|\hat{\theta}_T - \theta^0\right\| = O_P\left(\frac{1}{\lambda_{l,T}\sqrt{J_{l,T}}}\right).$$

Note that, in contrast with AR12, we got a result on rate of convergence without assuming anything about the differentiability of the moment conditions with respect to the parameters since our maintained assumption 4(ii) provides a Lipschitz condition that replaces the mean-value theorem argument in AR12.

3.2 Asymptotic normality

The result in theorem 2 assigns the slowest possible rate to all components of the structural parameters but, as shown in AR12, it is possible to achieve different rates of convergence in different directions in the parameter space. Following the procedure suggested in AR12, we discuss asymptotic normality of $\hat{\theta}_T$ with faster rates of convergence for some directions, paving the way for efficient inference.

This subsection must be seen as an extension of the section 4 of AR12 to the case of a number of moment conditions that may go to infinity with the sample size. Since smoothness of the moment conditions has not been assumed to characterize the rate of convergence, we need now to maintain a local identification assumption similar to assumption 3 in AR12. Note however that when the number of moments may go to infinity, the standard rank condition in assumption 6(iii) must be reinforced by assumption 6(iv) (that is redundant, up to the limit argument, when J_T is bounded).

Assumption 6 (Local identification): (i) $\rho_{j,T}(\theta)$ is continuously differentiable on the interior of Θ for all $j = 1, 2, \cdots$. (ii) θ^0 belongs to the interior of Θ. (iii) The $J_T \times p$ matrix $\frac{\partial \rho_T(\theta)}{\partial \theta'}$ has a full column rank p for all $\theta \in \Theta$. (iv) For all $i = 1, 2, \cdots, l$, $\lim_{T\to\infty} \frac{1}{J_{i,T}} \frac{\partial \rho_{iT}(\theta^0)'}{\partial \theta} \frac{\partial \rho_{iT}(\theta^0)}{\partial \theta'}$ exists and is positive definite.

Assumption 7 reinforces assumption 6 in the same way assumption 4 in AR12 reinforces their assumption 3.

Assumption 7: There exists non negative integers s_i for $i = 1, \cdots, l$ such that for all θ in

$int(\Theta)$,

$$rank[\Gamma_i(\theta^0)] = s_1 + s_2 + \cdots + s_i$$

where $\Gamma_i(\theta)$ is a $p \times (J_{1,T} + J_{2,T} + \cdots + J_{i,T})$ matrix such that $\sum_{i=1}^{l} s_i = p$ and

$$\Gamma_i(\theta)\Gamma_i(\theta)' = \lim_{T \longrightarrow \infty} \Gamma_{i,T}(\theta)\Gamma_{i,T}(\theta)',$$
$$\Gamma_{i,T}(\theta) = \left[\frac{1}{\sqrt{J_{1,T}}}\frac{\partial \rho_{1T}(\theta)'}{\partial \theta} \cdots \frac{1}{\sqrt{J_{i,T}}}\frac{\partial \rho_{iT}(\theta)'}{\partial \theta}\right].$$

Note that the first $J_{1,T}$ estimating equations ($\rho_{iT}(\theta)$) have a sample counterpart ($\psi_{iT}(\theta)$) converging at the rate $\lambda_{1,T}$. This $\lambda_{1,T}$ might not be the fastest rate appearing in Λ_T but they have the strongest identification strength considering $J_{1,T}$ together (recall that $\lambda_{1,T}\sqrt{J_{1,T}}$ is the fastest rate). In general, the rank of the Jacobian $\Gamma_1(\theta^0)$ is smaller than p and thus, the first set of estimating equations is not sufficient to identify θ. As stated in AR12, they only identify the s_1 directions in the space of parameters that correspond to col $\left[\Gamma_1(\theta^0)\right]$[2]. Similarly, col$\left[\Gamma_2(\theta^0)\right]$ only identifies the $(s_1 + s_2)$ directions in the parameter space that can be estimated at rate $\lambda_{2,T}\sqrt{J_{2,T}}$ at least. However, we know that the first s_1 directions can be estimated at a faster rate which invokes the need to disentangle those two rates for efficient estimation. If $s_1 + s_2 < p$, the same logic applies to $\Gamma_1(\theta^0), \cdots, \Gamma_l(\theta^0)$ until $s_1 + \cdots + s_l = p$.

In order to disentangle the various rates of convergence, we adapt the same reparameterization method proposed by AR12. We define a $p \times p$ non-singular matrix

$$R^0 = [R_1 R_2 \cdots R_l]$$

with $rank[R_l] = s_l$ and $rank[R_i R_{i+1} \cdots R_l] = s_i + s_{i+1} + \cdots + s_l$ such that $\frac{\partial \rho_{iT}(\theta^0)}{\partial \theta'} R_n = 0$ for $i, n = 1, 2, \cdots, l$ and $i < n$. This matrix R^0 is designed to separate the parameter space into the l number of subspaces such that each of them is associated with the different rates of convergence, $\lambda_{i,T}\sqrt{J_{i,T}}$. It enables us a new parameterization

$$\eta_T = \left[R^0\right]^{-1}\theta = \left[\eta_{1,T}' \ \eta_{2,T}' \ \cdots \ \eta_{l,T}'\right]', \quad \dim(\eta_{i,T}) = s_i \ \text{ for } \ i = 1, 2, \cdots, l,$$

and this helps disentangle the various rates of convergence. In order to see that, consider the estimating equations $\rho_T(\theta) = \rho_T(R^0\eta_T)$. Then $\partial \rho_T(R^0\eta_T)/\partial \eta_T' = \left[\partial \rho_T(R^0\theta)/\partial \theta'\right] R^0$, which is lower triangular for $\eta_T^0 = \left[R^0\right]^{-1}\theta^0$ by construction. We can see below (also shown in AR12 for $J_T < \infty$) that due to this lower triangularity we have, with some assumptions,

$$\lambda_{i,T}\sqrt{J_{i,T}}\left[\hat{\eta}_{i,T} - \eta_{i,T}^0\right] = O_p(1), \quad \text{for } \ i = 1, 2, \cdots, l,$$

[2]For any $m \times n$ matrix A, col$[A]$ represents the subspace of $\mathbb{R}^m$ generated by the column vectors of A.

where $\hat{\eta}_T = [\hat{\eta}'_{1,T} \ \hat{\eta}'_{2,T} \ \cdots \ \hat{\eta}'_{l,T}]' = \left[R^0\right]^{-1} \hat{\theta}_T$. Thus, each $\hat{\eta}_{i,T}$ displays a faster rate of convergence inherited from the moments $\psi_{iT}(\theta)$ than $\hat{\eta}_{j,T}$ for $i < j$.

Assumption 8 below is equivalent to assumption 5(ii) in AR12 except that we take into account that the norm of the Jacobian matrix may grow with T like the square root of the number of moment conditions.

Assumption 8: For each $i = 1, 2, \cdots, l$,

$$\frac{\partial \Phi_{iT}(\theta^0)}{\partial \theta'} = \sqrt{T}\left[\frac{\partial \bar{\psi}_{iT}(\theta^0)}{\partial \theta'} - \frac{\lambda_{i,T}}{\sqrt{T}}\frac{\partial \rho_{iT}(\theta^0)}{\partial \theta'}\right] = O_p(\sqrt{J_{i,T}}).$$

Assumption 9 provides a condition that prevents the estimation of $\eta_{i,T}$ from being contaminated by the estimation of $\eta_{j,T}$ that belongs to the parameter space with weaker identification strength when $i < j$. It is a sufficient condition to ensure

$$\frac{\sqrt{T}}{\lambda_{j,T}\sqrt{J_{j,T}}}\frac{\partial \bar{\psi}_{iT}(\tilde{\theta}_T)}{\partial \theta'}R_j \xrightarrow{p} 0, \text{ when } T \longrightarrow \infty \ \forall i < j. \tag{3.3}$$

where $\tilde{\theta}_T$ lies between $\hat{\theta}_T$ and θ^0 component by component. Although it is already shown in AR12 why we need this orthogonality condition for $J_T < \infty$, it is worth revisiting it.

First, consider the unfeasible minimum distance problem[3] for $J_T < \infty$:

$$\min_{\eta} \ \bar{\psi}_T(R^0\eta)' W_T \bar{\psi}_T(R^0\eta).$$

Then the GMM estimator $\hat{\eta}_T$ is the one that satisfies the following first order condition

$$R_T^{0\prime}\frac{\partial \bar{\psi}_T(R^0\hat{\eta}_T)'}{\partial \theta} W_T \bar{\psi}_T(R^0\hat{\eta}_T) = 0.$$

and the asymptotic distribution of $\hat{\theta}_T$ is derived using the first order Taylor expansion of $\sqrt{T}\bar{\psi}_T(R^0\hat{\eta}_T)$ at around η^0

$$\sqrt{T}\bar{\psi}_T(R^0\eta^0) + \sqrt{T}\frac{\partial \bar{\psi}_T(\tilde{\theta}_T)}{\partial \theta'}R^0(\hat{\eta}_T - \eta^0).$$

Then, for $i = 1, 2, \cdots, l$, the i-th group of this expansion is

$$\sqrt{T}\bar{\psi}_{iT}(R^0\eta^0) + \sum_{j=1}^{l}\frac{\sqrt{T}}{\lambda_{j,T}}\frac{\partial \bar{\psi}_{iT}(\tilde{\theta}_T)}{\partial \theta'}R_j^0\lambda_{j,T}(\hat{\eta}_{j,T} - \eta_j^0).$$

When $J_T < \infty$, $\lambda_{j,T}(\hat{\eta}_{j,T} - \eta_j^0) = O_p(1)$ and we need $\frac{\sqrt{T}}{\lambda_{j,T}}\frac{\partial \bar{\psi}_{iT}(\tilde{\theta}_T)}{\partial \theta'}R_j \xrightarrow{p} 0$ for $i < j$ to avoid the

[3]It is infeasible since R^0 is not known.

contamination. However, when $J_T \to \infty$ and $J_{i,T} \to \infty$ for some $i = 1, 2, \cdots, l$, the orthogonality condition should hold for the rescaled Jacobian with $\sqrt{J_{i,T}}$. Then, it looks reasonably intuitive that it is $\lambda_{i,T}\sqrt{J_{i,T}}(\hat{\eta}_{i,T} - \eta_j^0)$ that is $O_p(1)$, rather than with the convergence rate $\lambda_{i,T}$ as in the case with $J_T < \infty$.

Assumptions 9(ii) and (iii) are akin to assumption 6* in AR12. When J_T is finite, assumption 9(i) is equivalent to $\lambda_{i,T} \to \infty$ for all $i = 1, 2, \cdots, l$. However, when J_T grows with the sample size, it is not sufficient to have all moments are stronger than Stock and Wright (2000) sense but they have to be stronger. Note that for the orthogonality condition in (3.3) itself, it suffices to have

$$\lambda_{i,T} \longrightarrow \infty, \ \forall i = 1, 2, \cdots, l.$$

so that $J_T/\lambda_{l,T}\sqrt{J_{l,T}}$ is not restricted to converge to zero. However, assumption 9(i) is imposed to avoid the correlation between the Jacobians and the moments which causes asymptotic bias of the GMM estimator[4]. This condition is also maintained in HP such that $J_T/(Tc_T)^{1/2} = \sqrt{J_T}/\lambda_T \to 0$ to ensure $\lambda_T\sqrt{J_T}(\hat{\theta}_T - \theta^0)$ is asymptotically normal with zero mean.

Assumption 9: (i) $J_T/\lambda_{l,T}\sqrt{J_{l,T}} \to 0$ as $T \to \infty$. (ii) $\lambda_{1,T}\sqrt{J_{1,T}} = o(\lambda_{l,T}^2 J_{l,T})$. (iii) For $i = 1, 2, \cdots, l$, $k = 1, 2, \cdots, J_{iT}$, and $s = 1, 2, \cdots, p$, let $\frac{\partial \bar{\psi}_{iT,ks}(\theta)}{\partial \theta'}$ be the (k, s)-th element of $\frac{\partial \bar{\psi}_{iT}(\theta)}{\partial \theta'}$. Then $\frac{\sqrt{T}}{\lambda_{i,T}} \frac{\partial^2 \bar{\psi}_{iT,ks}(\theta)}{\partial \theta \partial \theta'}$ converges in probability to some well-defined matrix $H_{iks}(\theta)$ uniformly on Θ.

We define $\tilde{\Lambda}_T$ as a $p \times p$ matrix such that

$$\tilde{\Lambda}_T = \begin{bmatrix} \lambda_{1,T}\sqrt{J_{1,T}} Id_{s_1} & 0 & 0 & 0 \\ 0 & \lambda_{2,T}\sqrt{J_{2,T}} Id_{s_2} & 0 & 0 \\ ... & ... & ... & ... \\ 0 & 0 & 0 & \lambda_{l,T}\sqrt{J_{l,T}} Id_{s_l} \end{bmatrix}.$$

This matrix contains l number of rates associated with $s_1, s_2, \cdots, s_l$ directions of the parameter space. They are the proper rescaling rate of the sample Jacobian $\frac{\sqrt{T}\partial \psi_{iT}(R^0 \eta_T)}{\partial \eta_T'}$ for each $i = 1, 2, \cdots, l$ such that $\frac{\sqrt{T}\partial \psi_T(R^0 \eta_T)}{\partial \eta_T'}$ is asymptotically block-diagonal thanks to the reparameterization and the orthogonality condition. The i-th group of $\tilde{\Lambda}_T$ is in fact the rate of convergence of $\hat{\eta}_{i,T}$ for all $i = 1, 2, \cdots, p$, which improves as the number of moments increases.

Assumption 10 is akin to lemma 4.1 in AR12. While the proof of their lemma 4.1 is based on their assumptions 5 and 6*, it would be easy to use our assumptions 8 and 9 to adapt the argument of AR12, proving the validity of assumption 10 (hence a redundant assumption with maintained assumptions 8 and 9) in the case of a bounded number of moments. However, since we need to

[4]We still allow an asymptotic bias coming from misspecification by allowing $\|\Lambda_T \rho_T(\theta^0)\|$ to be O(1) rather than o(1).

control the norm of remainder terms in Taylor expansions while the dimension goes to infinity, we rather choose to maintain the high level assumption.

Assumption 10: Suppose we have θ_T such that $\left\|\theta_T - \theta^0\right\| = O_p\left(\frac{1}{\lambda_{l,T}\sqrt{J_{l,T}}}\right)$. Then

$$\left\|\sqrt{T}\frac{\partial\bar{\psi}_T(\theta_T)}{\partial\theta'}R^0\tilde{\Lambda}_T^{-1} - D_T^0\right\| = o_p\left(\frac{1}{\sqrt{J_T}}\right)$$

where D_T^0 is a $J_T \times p$ block-diagonal matrix with i-th diagonal blocks $\left[\frac{1}{\sqrt{J_{i,T}}}\frac{\partial\rho_{iT}(\theta^0)}{\partial\theta'}R_i\right]$ for all $i = 1, 2, \cdots, l$.

Even though we maintain assumption 10 as a high-level one, it is worth realizing that it is quite natural since each coefficient of the $J_T \times p$ matrix involved is intuitively of order $o_P\left(\frac{1}{\lambda_{l,T}\sqrt{J_T}}\right)$, that is, a fortiori (by virtue of assumption 11(i)) of order

$$o_P\left(\frac{1}{J_T}\right).$$

Thanks to the block diagonality property in assumption 10, we can now get differential rates of convergence to a Gaussian distribution for the reparameterized estimator $\hat{\eta}_T$. Using the usual first order approximation, we can derive the asymptotic distribution of the GMM estimator from this:

$$\tilde{\Lambda}_T^{-1}\left[R^0\right]^{-1}(\hat{\theta}_T - \theta^0) \approx \left[D_T^{0\prime}W_T D_T^0\right]^{-1} D_T^{0\prime}W_T\sqrt{T}\bar{\psi}_T(\theta^0).$$

Then the conditions given in assumption 11 ensures that $\tilde{\Lambda}_T^{-1}\left[R^0\right]^{-1}(\hat{\theta}_T - \theta^0)$ behaves well asymptotically. Assumption 11(ii) is natural in our setup since $D_T^{0\prime}$ is a properly scaled block-diagonal matrix and W_T is assumed to behave well asymptotically from the beginning of this paper. $D_T^{0\prime}W_T D_T^0$ is a $p \times p$ matrix that is bounded asymptotically since $\|W_T\| = O(1)$ and $D_T^{0\prime}D_T^0$ is a block-diagonal matrix with $s_i \times s_i$, $i = 1, \cdots, l$ blocks such that the i-th block matrix is

$$\frac{1}{J_{i,T}}(R_i^0)'\frac{\partial\rho_{iT}(\theta^0)'}{\partial\theta}\frac{\partial\rho_{iT}(\theta^0)}{\partial\theta'}R_i^0.$$

This is positive definite for all T by assumption 6 that the eigenvalues of $\frac{1}{J_{i,T}}\frac{\partial\rho_{iT}(\theta^0)'}{\partial\theta}\frac{\partial\rho_{iT}(\theta^0)}{\partial\theta'}$ bounded away from zero and infinity and R_i^0 is p.d. Assumption 11(ii) means that $\frac{1}{J_{i,T}}\frac{\partial\rho_{iT}(\theta^0)'}{\partial\theta}\frac{\partial\rho_{iT}(\theta^0)}{\partial\theta'}$ converges to a nonrandom $s_i \times s_i$ matrix for all $i = 1, \cdots, l$ which is an intuitive conditions since each element in $\frac{\partial\rho_{iT}(\theta^0)'}{\partial\theta}\frac{\partial\rho_{iT}(\theta^0)}{\partial\theta'}$ is a summation of $J_{i,T}$ terms. Assumption 11(i) jointly with assumption 11(ii) allows us to get a non-degenerate asymptotic normality result for GMM estimators after change of basis through the matrix R^0. Assumption 11(iii) and (iv) just

state Lyapunov sufficient conditions for central limit theorem applied to $\sqrt{T}\bar{\psi}_T\left(\theta^0\right)$.

Assumption 11: Let $S_T = E[\Phi_T(\theta^0)\Phi_T(\theta^0)']$. As $T \to \infty$, (i) $D_T^{0\prime} W_T S_T W_T D_T^0 \to V$ where V is a nonrandom $p \times p$ p.d. matrix, (ii) $D_T^{0\prime} W_T D_T^0 \to D$ where D is a $p \times p$ p.d. matrix, (iii) $\{\psi_{t,T}(\theta^0)\}$ is weakly dependent, (iv) $\frac{1}{T^2}\sum_{t=1}^{T} E\left[\left\|\psi_{t,T}(\theta^0) - E[\psi_{t,T}(\theta^0])\right\|^4\right] \to 0$.

Then with the asymptotic normality of moments, we have the asymptotic normality result of the GMM estimator where each s_i component has different rates of convergence. Note that we allow asymptotic bias that arises from local misspecification (assumption 5). Note that

$$(D_T^0)' W_T \sqrt{T}\bar{\psi}_T(\theta^0) = (D_T^0)' W_T \bar{\phi}_T(\theta^0) + (D_T^0)' W_T \Lambda_T \rho_T(\theta^0)$$

and we can get asymptotic normality of the GMM estimator with zero mean only if we can control $\Lambda_T \rho_T(\theta^0)$ (i.e. $\Lambda_T \rho_T(\theta^0) = o(1)$).

Theorem 3 (Asymptotic normality): Under the conditions of theorem 2 and assumptions 6 to 11,

$$\tilde{\Lambda}_T [R^0]^{-1}\left(\hat{\theta}_T - \theta^0\right) \xrightarrow{d} \mathcal{N}(D^{-1}B, D^{-1}VD^{-1}), \quad B = \underset{T\longrightarrow\infty}{lim}\,(D_T^0)' W_T \Lambda_T \rho_T(\theta^0),$$

where D and V are defined in assumption 11.

It is worth realizing that theorem 3 provides asymptotic normality of GMM estimators with a "textbook asymptotic variance matrix". Of course, this statement must be interpreted with caution since these formulas involve the matrix Λ_T whose diagonal terms diverge to infinity. This rescaling acknowledges that, with the reparameterization $\eta = [R^0]^{-1}\theta$, different components of η are estimated with different rates of convergence. However, this rescaling is immaterial for Wald inference (tests and confidence sets) since the Studentization will erase it. In other words, the practitioner may use standard formulas for GMM inference without paying attention to the underlying heterogeneity of rates of convergence. Moreover, examination of the asymptotic variance of the GMM estimator shows that with $W_T = S_T^{-1}$, not only one gets a kind of efficiency property of the GMM estimators but also standard asymptotic chi-square distributions for standard Wald test statistics.

We have chosen not to much emphasize these properties for two reasons. First, the efficiency claim should be qualified with heterogenous rates of convergence. Second, the allegedly efficient GMM estimator may not be feasible due to the need of a weighting matrix W_T such that:

$$P \lim_{T\longrightarrow\infty} \frac{1}{J_T}\left\|W_T - S_T^{-1}\right\| = 0$$

while the matrix S_T depends in general on the true unknown value θ^0 of the paper and has a

dimension J_T diverging to infinity with T. As already announced, the properties of two-step GMM or continuously-updated GMM in this context are beyond the scope of this paper. However, there are obviously very general circumstances under which the standard practice of Wald inference will remain valid.

4 Conclusion

We have provided a feasible asymptotic theory for GMM estimators that would justify standard Wald-type inference, in spite of a large number of moment conditions with heterogenous identification strengths. By this endeavor, we explicitly display the benefits of many moments such that the divergence to infinity of the number of moments at stake may improve the rate of convergence of GMM estimators (after a convenient rotation in the parameter space) when moments are weak or nearly-weak. Due to the heterogeneity of identification strengths, it may be the case that each coefficient of the vector of parameters is estimated with the slowest rate of convergence, because its estimator is contaminated by the noise coming from the weakest instruments. We show how to disentangle different directions in the parameter space with different identification strengths, using a convenient re-parameterization technique that enables us to define a new parameter as linear combinations of the structural parameter. With this, we identify directions with different rates of convergence. By doing so, we extend to nonlinear settings and reinforce the argument of CS: the consistent estimation depends importantly not only on the strength of instruments but also on the number of moment conditions. With heterogeneous identification strengths across instruments, we distinguish directions in the parameter space for which the rate of convergence is defined by the product of the "identification strength" (non-linear generalization of concentration parameter) in this direction times the square root of the number of moments with this strength. In other words, we shed some new light on the use of alternative asymptotics based on many (nearly) weak moment sequences like those of CS, HP, Newey and Windmeijer (2009) and AR12. Moreover, it is worth stressing that our context of (nearly) weak identification is more general than the most commonly used since, besides standard instruments weakness, we also consider cases where "(nearly) weak identification" is the result either of near-multicollinearity with a large number of regressors or of nonparametric rates of convergence. We also contribute to the literature on instruments weakness through a more intrinsic concept of (nearly) weakness, in the line of Antoine and Renault (2009) and AR12. Finally, we emphasize that by our strategy of identification of directions in the parameter space for which asymptotic normality of estimators is attained with the best possible rates, we pave the way for efficient inference. The price to pay for this achievement is that we consider a given sequence of weighting matrices, so that our asymptotic theory is not impacted by the fact that we have to estimate a variance matrix with increasing dimension. In this respect, we add some

rationale for Cochrane (2005)'s argument that there is some motivation to use pre-specified (non-efficient) weighting matrices for GMM. In our case, it is even not really giving up efficiency since it allows us to identify different efficient rates of convergence in different directions of the parameter space.

References

Ai, C. and X. Chen (2003): "Efficient estimation of models with conditional moment restrictions containing unknown functions," *Econometrica*, 71, 1795–1843.

Anatolyev, S. (2012): "Inference in regression models with many regressors," *Journal of Econometrics*, 170, 368–382.

Andrews, D. W. K. and X. Cheng (2012): "Estimation and inference with weak, semi-strong, and strong identification," *Econometrica*, 80, 2153–2211.

Antoine, B. and E. Renault (2009): "Efficient GMM with nearly-weak instruments," *The Econometrics Journal*, 12, S135–S171.

——— (2011): "Efficient inference with poor instruments: a general framework," *Handbook of Empirical Economics and Finance,* ed. by A. Ullah and D.E.A. Giles*,* CRC Press*,* chap. 2.

——— (2012): "Efficient minimum distance estimation with multiple rates of convergence," *Journal of Econometrics*, 170, 350–367.

——— (2019): "Testing identification strength," *Journal of Econometrics,* Forthcoming.

Caner, M. (2008): "Nearly-singular design in GMM and generalized empirical likelihood estimators," *Journal of Econometrics*, 144, 511–523.

Cattaneo, M. D., M. Jansson, and W. K. Newey (2018): "Alternative asymptotics and the partially linear model with many regressors," *Econometric Theory*, 32, 277–301.

Chao, J. C. and N. R. Swanson (2005): "Consistent estimation with a large number of weak instruments," *Econometrica*, 73, 1673–1692.

Cheng, X. (2015): "Robust inference in nonlinear models with mixed identification strength," *Journal of Econometrics*, 189, 207–228.

Cochrane, J. H. (2005): *Asset pricing*, Princeton University Press.

Frazier, D. T., E. Renault, L. Zhang, and X. Zhao (2019): "Weak Instruments Test in Discrete Choice Models," WP Monash University.

Gagliardini, P., C. Gourieroux, and E. Renault (2011): "Efficient derivative pricing by the extended method of moments," *Econometrica*, 79, 1181–1232.

Hahn, J. and G. Kuersteiner (2002): "Discontinuities of weak instrument limiting distributions," *Economics Letters*, 75, 325–331.

Hall, A. R. (2005): "Generalized Method of Moments, advanced texts in econometrics," Oxford University Press.

Han, C. and P. C. B. Phillips (2006): "GMM with many moment conditions," *Econometrica*, 74, 147–192.

Hansen, L. P. (1982): "Large sample properties of Generalized Method of Moments estimators," *Econometrica*, 50, 1029–1054.

Hausman, J. A., W. K. Newey, T. Woutersen, J. Chao, and N. R. Swanson (2012): "Instrumental variable estimation with heteroskedasticity and many instruments," *Quantitative Economics*, 3, 211–255.

Lewbel, A. (2007): "A local generalized method of moments estimator," *Economics Letters*, 94, 124–128.

Newey, W. K. and F. Windmeijer (2009): "Generalized Method of Moments with many weak moment conditions," *Econometrica*, 77, 687–719.

Staiger, G. and J. H. Stock (1997): "Instrumental variables regression with weak instruments," *Econometrica*, 65, 557–586.

Stock, J. H. and J. H. Wright (2000): "GMM with weak identification," *Econometrica*, 68, 1055–1096.